# Reading Time and Vocabulary Rating in the Japanese Language : Large-Scale Reading Time Data Collection Using Crowdsourcing

*Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 5178–5187
Marseille, 20-25 June 2022


# Reading Time and Vocabulary Rating in the Japanese Language: Large-Scale Reading Time Data Collection Using Crowdsourcing

**Masayuki Asahara**
National Institute for Japanese Language and Linguistics, Japan
Tokyo University of Foreign Studies
masayu-a@ninjal.ac.jp

**Abstract**

This study examined the effect of the differences in human vocabulary on reading time. This study conducted a word familiarity survey and applied a generalised linear mixed model to the participant ratings, assuming vocabulary to be a random effect of the participants. Following this, the participants took part in a self-paced reading task, and their reading times were recorded. The results clarified the effect of vocabulary differences on reading time.

**Keywords:** Japanese, Reading Time, Vocabulary Rating

## 1. Introduction

This study collected large-scale reading time data for the Japanese language using crowdsourcing. The English language, Natural Story Corpus (NSC) (Futrell et al., 2018) was built by recruiting people to participate in Amazon Mechanical Turk to collect reading times. In a similar experiment, this study recruited participants through Yahoo! Japan's crowdsourcing and utilised ibex[1] to collect large-scale reading time data using a self-paced reading method via a web browser. In addition, this study conducted a word familiarity rating experiment (Asahara, 2019) to evaluate word familiarity and the subject's vocabulary rating. The correlation between reading time and subject vocabulary size was estimated based on the collected reading time and vocabulary rating data. This study demonstrated a method for collecting reading times and presented the results of statistical analyses. The result shows that the vocabulary size are inversely correlated to the reading time.

## 2. Literature Review

This study surveyed datasets of reading time data from the corpus using eye-tracking and self-paced reading methods. The Potsdam Sentence Corpus (Kliegl et al., 2004; Kliegl et al., 2006) is a reading time corpus of 144 German sentences (1138 words) read by 222 participants. The Dundee Eye-Tracking Corpus (Kennedy and Pynte, 2005) (Kennedy, 2003) contains reading times for English and French newspaper editorials from 10 native speakers for each language, recorded using eye-tracking equipment. The English version of the Dundee Eye-Tracking Corpus is composed of 20 editorial articles with 51,501 words. The Dutch DEMONIC database (Kuperman et al., 2010) consists of 224 Dutch sentences with reading times of 55 participants. The UCL Corpus (Frank et al., 2013) consists of 361 English sentences drawn from amateur novels with self-paced reading and eye-tracking data. The sentences are presented without contexts. The Potsdam-Allahabad Hindi Eyetracking Corpus (Husain et al., 2015) consists of 153 Hindi-Urdu treebank sentences with eye-tracking data from 30 participants. The Ghent Eye-Tracking Corpus corpus (Cop et al., 2017) (Cop et al., 2007) contains monolingual and bilingual (L1 and L2) corpuses with eye-tracking data of participants reading a complete novel *The Mysterious Affair at Styles* by Agatha Christie. The Natural Stories corpus (Futrell et al., 2018; Futrell et al., 2021) (Futrell et al., 2021) consists of 10 stories of approximately 1,000 words each (10,245 lexical word tokens) with self-paced reading data from 181 English speakers recruited through Amazon Mechanical Turks. The Provo Corpus (Luke and Christianson, 2018) (Luke, 2017) is an eye-tracking corpus of 84 American English speakers, which consists of 2,689 words. The Russian Sentence Corpus (Laurinavichyute et al., 2019) contains 144 naturally occurring sentences extracted from the Russian National Corpus. Full sentences were presented on the screen to monolingual Russian speakers. The Potsdam Textbook Corpus (Jäger et al., 2021) contains 12 short passages of 158 words on average from college-level biology and physics textbooks, which are read by German native speakers, including experts and laypersons. The Beijing Sentence Corpus (Pan et al., 2021) includes eye-tracking data from 60 Chinese participants.

Although large-scale reading time databases have been compiled in various languages, Japanese reading time data are limited to BCCWJ-EyeTrack (Asahara et al., 2016). BCCWJ-EyeTrack consists of 20 newspaper articles in the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) with reading data of 24 participants based on eye-tracking and self-paced reading. BCCWJ contains rich annotation data, such as syntactic dependency structures (Asahara and Matsumoto, 2016), information struc-

[1] https://github.com/addrummond/ibexfarm

tures (Miyauchi et al., 2017), syntactic and semantic categories (Kato et al., 2018), and co-reference information, including exophora and clause boundaries (Matsumoto et al., 2018). Annotation labels and reading time were compared (Asahara, 2017; Asahara and Kato, 2017; Asahara, 2018b; Asahara, 2018a). However, the text amount and number of participants were limited. This study aimed to increase the reliability of the results by increasing the amount of data. Eye-tracking experiments could not be performed in the laboratory due to the COVID-19 pandemic; therefore, large-scale self-paced reading time data were collected via crowdsourcing to precede the Natural Stories Corpus (Futrell et al., 2018).

Following this, datasets of word familiarity ratings were surveyed. The NTT Database Series: Lexical Properties of Japanese (*Nihongo-no Goitokusei*) (NTT Communication Science Laboratories, 1999 2008) is the world's largest database that examines lexical features from a variety of perspectives, aiming to clarify Japanese language functions. In addition, it contains subjective data, such as word familiarity, orthography-type appropriateness, word accent appropriateness, kanji familiarity, complexity, reading appropriateness, and word mental image, and objective data based on the frequency of vocabulary as it appears in newspapers. The Word Familiarity Database (Heisei Era Version) (Amano and Kondo, 1998)(Amano and Kondo, 1999; Amano and Kondo, 2008) is an advanced lexical database on the familiarity of vocabulary. The Word Familiarity Database (Reiwa Era Version) (NTT Communication Science Laboratories, 2021) was created, as it was noted that people's perceptions of vocabulary changed since the first survey. The word mental image characteristic database collected information on the 'ease of sensory imagery of semantic content' for written and spoken stimuli (Sakuma et al., 2008). In addition, NINJAL has been continuously working on the estimation of word familiarity for the WLSP (Asahara, 2019) and has published several lexical tables.

# 3. Collecting Reading Time Data

## 3.1. Methods for Collecting Reading Time Data

### 3.1.1. Stimulus Sentences

Table 1: The statistics of the stimulus sentences

| registers | | samples | sentences | phrases |
|---|---|---|---|---|
| OW | White paper | 1 | 36 | 462 |
| OT | Textbooks | 38 | 9,521 | 50,606 |
| (avg. per sample) | | | 250.6 | 1331.7 |
| PB | Books | 83 | 10,075 | 84,736 |
| (avg. per sample) | | | 121.4 | 1,020.9 |

This study used white paper (OW), textbooks (OT), and book (PB) samples[2] from the BCCWJ (Maekawa et al., 2014) for stimulus sentences. OW is one copyright-free sample in BCCWJ. For OT, 38 samples (17 elementary, 9 middle, and 12 high school) of Japanese language classes (contemporary writing) were used. Reading time data on these stimulus texts were collected, assuming that anyone who had received Japanese language education in Japan had read them in the past. This study assumed that it would be used as contrast data for readability evaluation in Japanese language education after obtaining data on people who received Japanese language education overseas or learners of the Japanese language. For PB, 83 samples of BCCWJ core data were used. The study was conducted based on various annotations, such as dependency on these data. It was assumed that the data would be used for commercial purposes, limited by the reading time data of PN. Table 1 shows the number of samples, sentences, and phrases in the stimulus sentences (for the number of recruited participants, see the next section). For comparison, the Natural Stories Corpus sampled approximately 1,000 words. The present data averaged over 1,000 characters as OW, OT, and PB samples in BCCWJ. For each stimulus sentence, two questions with 'yes' or 'no' answers were established to check whether the participants read the content correctly.

[2]One sample can be regarded as one document.

### 3.1.2. Self-Paced Reading Method

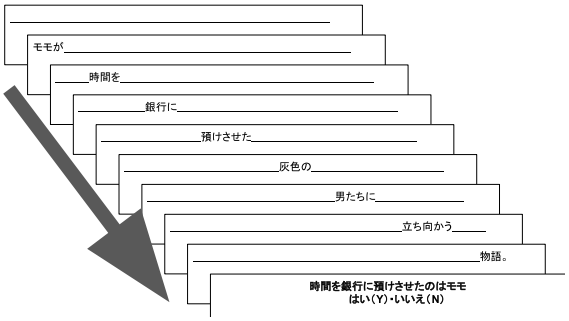



Figure 1: The self-paced reading method

The self-paced reading method used moving windows to set up an environment that moves the line of sight to measure reading times based on the display time of partially presented words and phrases. Figure 1 shows an example. The phrase was displayed sequentially each time the space key was pressed, and the reading time was measured by recording the time interval in which the space key was pressed in milliseconds. Comprehension questions that required the participants to answer 'yes' or 'no' were presented at the end of each passage. Ibexfarm was used to conduct the experiment using a self-paced reading test on a browser.

### 3.1.3. Participants

Participants were recruited through Yahoo! crowdsourcing. In October 2020, a word familiarity rating

Table 2: The statistics of the participants

| registers | | participants per sample |
|---|---|---|
| OW | White paper | 425 |
| OT | Textbooks | 200 |
| PB | Books | 200 |

experiment was conducted (Asahara, 2019) in section 4.1. This study recruited 2,092 people with a large variance in word-familiarity responses. This research was conducted as a preliminary experiment, as eligible participants could be selected in advance, and the vocabulary of each participant could be estimated when estimating word familiarity. For OW, this study recruited 500 people on a trial basis, and 427 people participated. For OTs and PBs, this study recruited 200 people.[3] This study recruited 100 participants who answered 'yes' to content-checking questions and 100 participants who answered 'no'. The number of participants are presented in Table 2.

## 3.2. Organising the Reading Time Data

### 3.2.1. Eliminating Inappropriate Data

Table 3: Elimination of inappropriate samples

| Description | OW | OT | PB |
|---|---|---|---|
| All samples | 425 | 7,453 | 16,553 |
| (1) disagree | 54 | 278 | 199 |
| (2) multiple time | 0 | 46 | 64 |
| (3) <150ms or 2,000ms< | 15 | 1,439 | 2,875 |
| (4) incorrect answer | 48 | 825 | 2,090 |
| Appropriate samples | 308 | 4,865 | 11,325 |

As the data in this study were collected online, it was assumed that data quality was worse than data collected in person. Therefore, inappropriate data were eliminated, including (1) data of respondents who did not agree with the method of handling the experimental data and the reward payment method at the start of the experiment, (2) repeated submissions by the same participant, with only the first submission retained, (3) responses with an average reading time per sample below 150 ms or above 2,000 ms, and (4) data of respondents who answers the 'yes' or 'no' question incorrectly.

Table 4: Elimination of inappropriate data points

| Description | OW | OT | PB |
|---|---|---|---|
| Appropriate samples | 140,449 | 6,114,814 | 11,309,783 |
| (5) <100 or 3,000< | 3,653 | 409,916 | 540,403 |
| Objects of analysis | 136,797 | 5,704,898 | 10,769,380 |

Inappropriate data for each data point were eliminated. The 'Appropriate samples' row in Table 4 shows the number of data point contained in the appropriate sample. An additional five points that were below 100 ms or over 3,000 ms[4] were eliminated. Consequently, the number of cases shown in 'Objects of analysis' in the table was considered appropriate data points.

### 3.2.2. Format of the Reading Time Data

Table 5 shows the format of the experimental data. The data are in the TSV form. BCCWJ_Sample_ID and BCCWJ_start are the location information in BCCWJ. This information enabled the comparison of morphological information and various annotation data in BCCWJ. SPR_sentence_ID and SPR_bunsetsu_ID are the sentence ID and phrase ID in the experiment, which indicated the order of presentation in the experiment. SPR_surface is a surface form. OT and PB are masked so that only the number of characters (SPR_word_length) can be seen. SPR_reading_time is the reading time to be analysed. In addition, a logarithmic reading time (SPR_log_reading_time) was generated during the analysis. SPR_instructiontime is the time when the first instruction was given. Reading instructions were skipped in the second and subsequent experiments. SPR_QA_question is a 'yes' or 'no' question for checking during the experiment, and SPR_QA_answer is the correct answer. SPR_QA_correct is a flag indicating whether the participant answered the 'yes' or 'no' question correctly. The published data comprised only correct answers. SPR_QA_qatime is the time required for the participant to answer the 'yes' or 'no' question. SPR_subj_ID is the participant ID. SPR_averageRT is the average reading time per sample during the experiment. SPR_timestamp indicates the time when the experiment was performed. SPR_trial indicates the number of times the reading time of the genre was measured. SPR_control is the method of presenting information during the experiment (whether there was a space between phrases). To examine the impact of dependency on reading time, the information of BCCWJ-DepPara (Asahara and Matsumoto, 2016) was added to the core data OW/PB. As BCCWJ-DepPara defines a sentence boundary different from BCCWJ, there was a discrepancy with the phrase ID when presented (DepPara_bid, DepPara_depid). The number of dependencies (DepPara_depnum) was based on the BCCWJ-DepPara standard. For OT, textbooks of school type (BCCWJ_OT_school_type, OT01: elementary, OT02: middle, OT03: high school) were set as the object of analysis.

## 3.3. Preliminary Analysis of the Reading Time Data

### 3.3.1. Analytical Method

This section presents the preliminary analysis results of a frequency-based analytical method (Baayen, 2008; Vasishth et al., 2021). Reading times were exam-

[3] In some samples, 200 or more people were recruited due to a mistake made in the experimental set-up in OT and PB.

[4] As a criterion, we referred to the Natural Stories Corpus(Futrell et al., 2018), which removed data points below 100 ms or above 3,000 ms.

Table 5: Data format

| column name | description | OW | OT | PB |
|---|---|---|---|---|
| BCCWJ_Sample_ID | sample ID in BCCWJ | ✓ | ✓ | ✓ |
| BCCWJ_start | position in BCCWJ | ✓ | ✓ | ✓ |
| SPR_sentence_ID | sentence ID | ✓ | ✓ | ✓ |
| SPR_bunsetsu_ID | phrase ID | ✓ | ✓ | ✓ |
| SPR_surface | surface form | ✓ | masked | masked |
| SPR_word_length | number of characters | ✓ | ✓ | ✓ |
| SPR_sentence | sentence | ✓ | masked | masked |
| SPR_reading_time | reading time | ✓ | ✓ | ✓ |
| SPR_log_reading_time | logarithm of reading time | ✓ | ✓ | ✓ |
| SPR_instruction_time | instruction time | ✓ | ✓ | ✓ |
| SPR_QA_question | YES/NO question | ✓ | ✓ | ✓ |
| SPR_QA_answer | correct answer of YES/NO question | ✓ | ✓ | ✓ |
| SPR_QA_correct | correct or not in YES/NO question | ✓ | ✓ | ✓ |
| SPR_QA_qa_time | time of YES/NO question | ✓ | ✓ | ✓ |
| SPR_subj_ID | participant ID | ✓ | ✓ | ✓ |
| SPR_averageRT | average reading time in a sample | ✓ | ✓ | ✓ |
| SPR_timestamp | timestamp of experiment | ✓ | ✓ | ✓ |
| SPR_trial | trial number per subject | N/A | ✓ | ✓ |
| SPR_control | presentation method | ✓ | ✓ | ✓ |
| DepPara_bid | phrase id in treebank | ✓ | N/A | ✓ |
| DepPara_depid | dependent id in treebank | ✓ | N/A | ✓ |
| DepPara_depnum | the number of dependents | ✓ | N/A | ✓ |
| BCCWJ_OT_school_type | school types for textbooks | N/A | ✓ | N/A |

ined using a generalised linear mixed model (R (R Core Team, 2020), lme4 (Bates et al., 2015), and stargazer (Hlavac, 2018)). SPR_sentence_ID (experimental sentence ID) and SPR_bunsetsu_ID (experimental phrase ID), which present information of the presentation order, and SPR_word_length, which is the number of words of surface form, were used as fixed effects. For OW and PB, the number of dependencies of the phrase, DepPara_depnum, were considered. For OT, school type, SPR_OT_school_type, was considered. For OT and PB, the trial order, SPR_trial, was modelled as a fixed effect when the same participant read multiple samples. This study considered SPR_subj_ID (participant ID) as a random effect to model individual differences between participants. For OT/PB, BCCWJ_Sample_ID (sample ID of BCCWJ) was used as a random effect to model individual differences between samples. The analytical formula is as follows:

$$\begin{aligned} \texttt{SPR\_reading\_time} \sim\ & \texttt{SPR\_sentence\_ID} \\ & + \texttt{SPR\_bunsetsu\_ID} + \texttt{SPR\_word\_length} \\ & + \texttt{SPR\_trial} + \texttt{DepPara\_depnum} \\ & + \texttt{BCCWJ\_OT\_school\_type} + \\ & + (1|\texttt{SPR\_subj\_ID\_factor}) \\ & + (1|\texttt{BCCWJ\_Sample\_ID}). \end{aligned} \quad (1)$$

Data points with values outside 3SD were eliminated and the analysis was conducted again.

### 3.3.2. Results

Table 6 and 7 shows the analysis results, including the estimates for the fixed effects with the presence or absence of significant differences. Values in parentheses are standard deviation. The presentation order (SPR_sentence_ID, SPR_bunsetsu_ID) in the samples tended to have a shorter reading time with any register. There was also a tendency for the reading times to increase as the length of words increased (SPR_word_length). The correlation between OW and PB and the number of dependencies of the phrase (DepPara_depnum) was examined. In both results, reading times tended to be shorter for phrases with a large number of dependencies. For OT and PB, the trial order when the same participant read multiple samples, SPR_trial, was considered as a fixed effect. For OT, the reading times decreased as the number of trials increased; however, for PB, the reading times increased as the number of trials increased. For OT, school type, SPR_OT_school_type, was considered. Participants tended to spend less time reading high school textbooks than elementary school textbooks. For OT, it was easy to proceed in the order of elementary school $\rightarrow$ middle school $\rightarrow$ high school on selection screens. There was a correlation between SPR_trial and the elementary, middle, and high school types. This may have led to discrepancies between the registers in the effect of the number of trials in reading times. The participants tried OW, OT, and PB in a specified order. In the first experiment, the reading times tended to be longer than those of the other registers. OT includes textbooks likely to be encountered during Japanese language education in Japan. However, due to the influence of the presentation order, the reading times tended to be longer than that of PB. For books, comparisons were made between genres using the Nippon Decimal Clas-

Table 6: Preliminary analysis results of reading time by GLMM

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | SPR_reading_time | | | | | |
| | OW (white paper) | | OT (textbook) | | PB (book) | |
| SPR_sentence_ID | $-6.042^{***}$ | (0.048) | $-0.125^{***}$ | (0.0004) | $-0.143^{***}$ | (0.001) |
| SPR_bunsetsu_ID | $-1.477^{***}$ | (0.046) | $-2.047^{***}$ | (0.011) | $-0.856^{***}$ | (0.006) |
| SPR_word_length | $24.311^{***}$ | (0.160) | $5.113^{***}$ | (0.021) | $6.705^{***}$ | (0.014) |
| DepPara_depnum | $-15.048^{***}$ | (0.555) | | | $-5.225^{***}$ | (0.033) |
| SPR_trial | | | $-0.760^{***}$ | (0.005) | $0.379^{***}$ | (0.006) |
| BCCWJ_OT_school_typeOT02 | | | $-7.721$ | (8.898) | | |
| BCCWJ_OT_school_typeOT03 | | | $-25.228^{***}$ | (8.352) | | |
| Constant | $540.050^{***}$ | (11.951) | $361.566^{***}$ | (7.155) | $306.887^{***}$ | (5.321) |
| data points | 133,806 | | 5,617,794 | | 10,701,504 | |
| elimination rate (outside 3SD) | 1,726 | (0.0126) | 87,103 | (0.0152) | 168,671 | (0.0155) |
| log-likelihood | $-897,781.800$ | | $-34,071,483.000$ | | $-64,679,004.000$ | |
| *Note:* | | | | | | $^{***}$p<0.01 |

Table 7: Preliminary analysis results of logarithm reading time by GLMM

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | SPR_log_reading_time | | | | | |
| | OW (white paper) | | OT (textbook) | | PB (book) | |
| SPR_sentence_ID | $-0.012^{***}$ | (0.0001) | $-0.0004^{***}$ | (0.00000) | $-0.0004^{***}$ | (0.00000) |
| SPR_bunsetsu_ID | $-0.003^{***}$ | (0.0001) | $-0.006^{***}$ | (0.00003) | $-0.003^{***}$ | (0.00002) |
| SPR_word_length | $0.036^{***}$ | (0.0002) | $0.010^{***}$ | (0.0001) | $0.014^{***}$ | (0.00004) |
| DepPara_depnum | $-0.022^{***}$ | (0.001) | | | $-0.012^{***}$ | (0.0001) |
| SPR_trial | | | $-0.002^{***}$ | (0.00001) | $0.001^{***}$ | (0.00002) |
| BCCWJ_OT_school_typeOT02 | | | $-0.025$ | (0.024) | | |
| BCCWJ_OT_school_typeOT03 | | | $-0.078^{*}$ | (0.022) | | |
| Constant | $6.216^{***}$ | (0.023) | $5.826^{***}$ | (0.020) | $5.664^{***}$ | (0.016) |
| data points | 135,070 | | 5,623,067 | | 10,707,218 | |
| elimination rate (outside 3SD) | 1,726 | (0.0126) | 81,830 | (0.0143) | 162,957 | (0.0150) |
| log-likelihood | $-38,248.550$ | | $-630,606.800$ | | $-780,846.500$ | |
| *Note:* | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

sification and other methods.

## 4. Reading Time with Vocabulary Rating

### 4.1. Building Vocabulary Rating Data

This study quantified the bias of the participants obtained during the collection of word familiarity data (Asahara, 2019) and used it for vocabulary data. Specifically, data on the extent to which participants knew, wrote, read, spoke, and heard the target word were collected. The list included 96,557 target words taken from the WLSP. Voice data (oral pronunciations) for the lexical entries were not included; however, speech and hearing were considered as two of the following five perspectives:

**KNOW:** How much do you know about the target word?

**WRITE:** How often do you write the target word?

**READ:** How often do you read the target word?

**SPEAK:** How often do you speak the target word?

**LISTEN:** How often do you hear the target word?

In this design, the judgments were split between character-based (WRITE and READ) and voice-based (SPEAK and LISTEN) judgments, and between production (WRITE and SPEAK) and reception (READ and LISTEN) judgments. The participants gave five ratings for each factor, ranging from 5 (well-known/often used) to 1 (little known/rarely used).

The collected data were modelled in the following ways: word familiarity as a random effect of 84,114 surface forms[5] and vocabulary as a random effect of 6,732 participants using a Bayesian linear mixed model. The graphical model used to estimate the ratings is shown in Figure 4: $N_{word}$ is the number of words (surface forms), $N_{subj}$ is the number of participants, Index $i : 1 \ldots N_{word}$ is the index of words, index $j : 1 \ldots N_{subj}$ is the index of participants, and $y^{(i)(j)}$ is the rating of KNOW, WRITE, READ, SPEAK, LISTEN. $y$ was generated by a normal distribution with $\mu^{(i)(j)}$ and $\sigma$, as follows:

$$y^{(i)(j)} \sim Normal(\mu^{(i)(j)}, \sigma).$$

$\sigma$ is a hyper-parameter of the standard deviation and

[5] 96,557 words include several homonyms and consist of 84,114 surface forms.

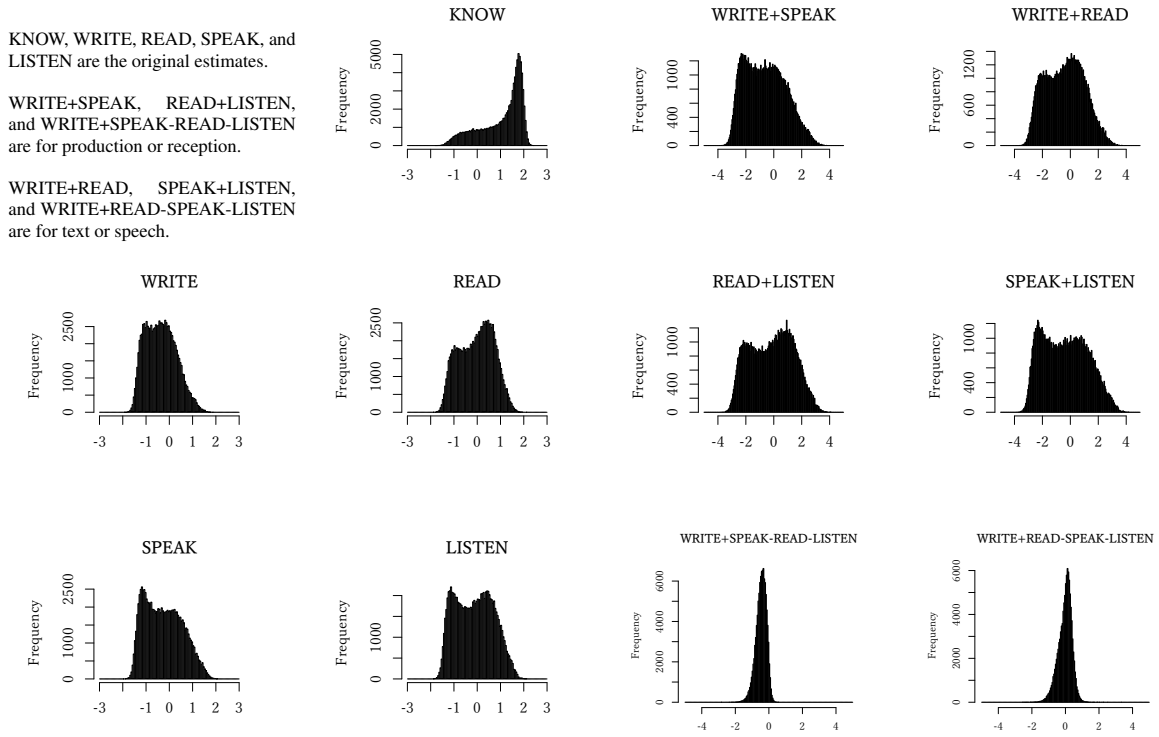



Figure 2: Results of word familiarity rating

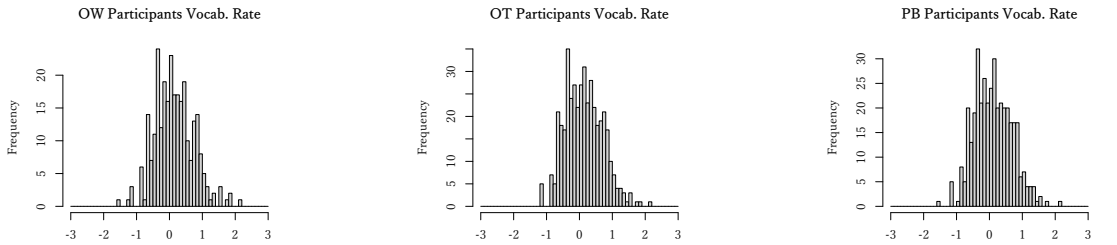



Figure 3: Results of vocabulary rating

$\mu^{(i)(j)}$ is a linear formula of slopes $\gamma^{(i)}_{subj}$, slopes $\gamma^{(i)}_{word}$, and an intercept $\alpha$:

$$\mu^{(i)(j)=\alpha+\gamma^{(i)}_{word}+\gamma^{(j)}_{subj}}.$$

The slopes were modelled by a normal distribution with the hyper-parameters of $\mu_{word}$, $\sigma_{word}$, $\mu_{subj}$, and $\sigma_{subj}$ (means and standard deviations):

$$\gamma^{(i)}_{word} \sim Normal(\mu_{word}, \sigma_{word}),$$

$$\gamma^{(j)}_{subj} \sim Normal(\mu_{subj}, \sigma_{subj}).$$

The word familiarity rates were composed by $\gamma^{(i)}_{word}$. The biases of subject participants were modelled by $\gamma^{(j)}_{subj}$.

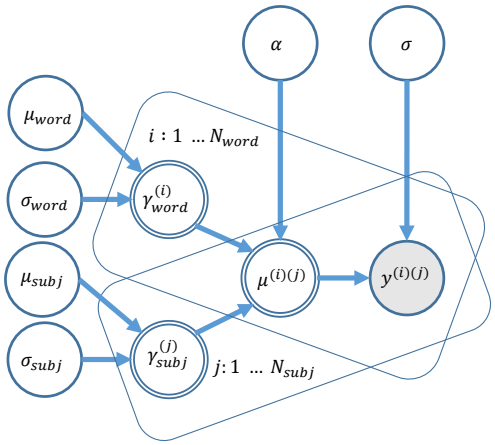
This study used a normal distribution with a standard deviation of 1.0 for words and 0.5 for the participants.

Figure 2 shows the estimated word familiarities. Figure 3 shows the distribution of the vocabulary of those who participated in the collection of reading time data. The frequentist model for the word familiarity rate estimation could not be constructed due to the number of parameters in the environment.

The data is available at `https://github.com/masayu-a/WLSP-familiarity`.

## 4.2. Reading Time with Vocabulary Rating

Data of the participants who participated in the word familiarity survey, including over 200 answers, and over five samples of the reading time experiment were analysed. Table 8 shows the statistics of the participants, and Table 9 shows the data with vocabulary ratings.




$$\gamma^{(i)}_{word} \sim Normal(\mu_{word}, \sigma_{word})$$
$$\gamma^{(j)}_{subj} \sim Normal(\mu_{subj}, \sigma_{subj})$$

$$\mu^{(i)(j)} = \alpha + \gamma^{(i)}_{word} + \gamma^{(j)}_{subj}$$
$$y^{(i)(j)} \sim Normal(\mu^{(i)(j)}, \sigma)$$

Figure 4: Graphical model for the Ratings

Table 8: The statistics of the participants (more than 200 answers in the word familiarity survey and more than five samples of reading time experiment)

| register | | participants (number of individuals) |
|---|---|---|
| OW | white paper | 277 |
| OT | textbooks | 422 |
| PB | books | 388 |

Table 9: The data with vocabulary rating results

| samples × subj. | OW | OT | PB |
|---|---|---|---|
| Data in section 3 | 308 | 4,865 | 11,325 |
| w/ vocab. rating | 277 | 4,685 | 10,932 |
| data points | OW | OT | PB |
| Data in section 3 | 136,797 | 5,704,898 | 10,769,380 |
| w/ vocab. rating | 124,502 | 5,490,977 | 10,484,300 |

### 4.3. Statistical Analyses with Vocabulary Rating

Statistical analyses were performed based on whether significant differences in each fixed effect could be identified in a linear mixed model for reading time and logarithmic reading times, as in Section 3.3. Presentation order, word length, trial order, dependency, and **vocabulary rating of the participants** were used as fixed effects, with participants and sample ID as random effects. Table 10 shows the results of the model estimation, with re-estimation after excluding data points outside three SDs. Table 11 shows the analysis of the frequency-based reading times using a frequency-based linear mixed model, and Table 12 in the Appendix shows the analysis of a similar model using the Bayesian linear mixed model (lognormal model).

First, for the order of presentation within the same sample, reading times to obtain context were shorter as the experiment progressed. Longer word length took longer to be recognised. The order of trials tended to be shorter in the textbooks as the participants became more familiar with it. The order of trials tended to be longer during the book task. The reason for this may be the fatigue caused by repeated implementation. A greater number of dependencies was associated with shorter reading times due to predictability. These results were similar to the preliminary analysis presented in Section 3.3.2.

The results showed that reading times tended to be shorter in the group with a larger vocabulary. The impact of vocabulary in terms of significant differences was $p < 0.1$ for OT (textbooks), $p < 0.05$ for PB (books), and $p < 0.01$ for OW (white papers). The textbooks were samples from elementary, middle, and high school Japanese textbooks, which the participants likely read at least once. Participants may generally be more familiar with books than white papers. This degree of familiarity with registers may have been reflected in the significant difference in the effect of vocabulary. In other words, it is possible that the difference in the influence of vocabulary was small, as the textbooks were familiar to the participants, and that the difference in the influence of vocabulary was large, as the white paper was unfamiliar.

## 5. Conclusion

This study presented a large-scale reading time data set constructed using crowdsourcing. A large-scale reading time dataset was constructed in a short period and at a low cost using ibexfarm, an experimental environment based on the self-paced reading method running in a browser to recruit participants through Yahoo! crowdsourcing. The registers of the stimuli texts were white papers, textbooks, and books. Due to the COVID-19 pandemic, eye-tracking experiments could not be conducted. While it was not possible to conduct the experiment face-to-face, a method of conducting a large-scale online survey to record reading times

Table 10: Analysis results of reading time with vocabulary rating by GLMM

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | SPR_reading_time | | | | | |
| | OW (white paper) | | OT (textbooks) | | PB (books) | |
| SPR_sentence_ID | $-6.087^{***}$ | (0.051) | $-0.127^{***}$ | (0.0004) | $-0.142^{***}$ | (0.001) |
| SPR_bunsetsu_ID | $-1.501^{***}$ | (0.049) | $-2.046^{***}$ | (0.011) | $-0.856^{***}$ | (0.006) |
| SPR_word_length | $24.820^{***}$ | (0.170) | $5.170^{***}$ | (0.021) | $6.798^{***}$ | (0.015) |
| SPR_trial | | | $-0.757^{***}$ | (0.005) | $0.382^{***}$ | (0.006) |
| DepPara_depnum | $-15.310^{***}$ | (0.591) | | | $-5.258^{***}$ | (0.034) |
| WFR_subj_rate (Vocab) | $-81.239^{***}$ | (21.227) | $-16.169^{*}$ | (9.087) | $-18.405^{**}$ | (8.731) |
| Constant | $558.984^{***}$ | (12.936) | $353.723^{***}$ | (6.548) | $306.631^{***}$ | (5.425) |
| data points | 121,769 | | 5,407,252 | | 10,321,560 | |
| elimination rate (outside 3SD) | 2,732 | (0.0219) | 83,724 | (0.0152) | 162,740 | (0.0155) |
| log-likelihood | −818,815.100 | | −32,796,021.000 | | −62,393,234.000 | |
| *Note:* | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 11: Analysis results of logarithm reading time with vocabulary rating by GLMM

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | SPR_log_reading_time | | | | | |
| | OW (white paper) | | OT (textbooks) | | PB (books) | |
| SPR_sentence_ID | $-0.012^{***}$ | (0.0001) | $-0.0004^{***}$ | (0.00000) | $-0.0004^{***}$ | (0.00000) |
| SPR_bunsetsu_ID | $-0.003^{***}$ | (0.0001) | $-0.006^{***}$ | (0.00003) | $-0.003^{***}$ | (0.00002) |
| SPR_word_length | $0.036^{***}$ | (0.0002) | $0.011^{***}$ | (0.0001) | $0.014^{***}$ | (0.00004) |
| SPR_trial | | | $-0.002^{***}$ | (0.00001) | $0.001^{***}$ | (0.00002) |
| DepPara_depnum | $-0.022^{***}$ | (0.001) | | | $-0.012^{***}$ | (0.0001) |
| WFR_subj_rate (Vocab) | $-0.180^{***}$ | (0.040) | $-0.052^{**}$ | (0.025) | $-0.052^{**}$ | (0.026) |
| Constant | $6.255^{***}$ | (0.025) | $5.826^{***}$ | (0.020) | $5.664^{***}$ | (0.016) |
| data points | 135,070 | | 5,412,398 | | 10,327,584 | |
| elimination rate (outside 3SD) | 1,559 | (0.0125) | 78,578 | (0.0143) | 156,716 | (0.0149) |
| log-likelihood | −38,816.700 | | −598,180.400 | | −743,585.500 | |
| *Note:* | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

and vocabulary in Japanese was developed. Before the reading time experiments, word familiarity rate experiments were conducted with the same participants. The effect of vocabulary on reading times in Japanese was investigated by conducting a large-scale survey with multi-register materials. The results of the study confirmed that the group with a large vocabulary had shorter reading times.

The data without the original textdata is available at https://github.com/masayu-a/BCCWJ-SPR2[6]. Further analyses will be conducted by comparing the data with various annotations of BCCWJ. The Bayesian analysis will also be conducted. In addition, we hope to digitise the process of acquiring language reading ability by collecting the reading times of L1 and L2 learners.

## 6. Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work has been partially supported by JSPS KAKENHI (Grant Number JP18H05521 and JP22H00663) and is NINJAL project 'Evidence-based Computational Psycholinguistics Using Annotation Data'.

[6]The original text is available at https://clrd.ninjal.ac.jp/bccwj/en/.

## 7. Appendix: Results of Bayesian Linear Mixed Model

Table 12: Analysis results of reading time with vocabulary rating by BLMM of lognormal

| OW(white paper) | mean | se_mean | sd |
|---|---|---|---|
| $\alpha$ intercept | 6.217 | 0.064 | 0.123 |
| $\beta_{\text{sentid}}$ | -0.012 | 0.000 | 0.002 |
| $\beta_{\text{bid}}$ | -0.003 | 0.000 | 0.001 |
| $\beta_{\text{length}}$ | 0.037 | 0.000 | 0.006 |
| $\beta_{\text{dependency}}$ | -0.024 | 0.001 | 0.009 |
| $\beta_{\text{subjrate}}$ | -0.118 | 0.094 | 0.116 |
| $\sigma$ | 0.988 | 0.075 | 0.097 |
| $\sigma_{\text{subj}}$ | 2.603 | 1.020 | 1.254 |

A Bayesian linear mixed model evaluation (Sorensen et al., 2016) by rstan was conducted. The following lognormal model was used:

```
model {
  real mu;
```

```
  // prior
  gamma_subj ~ normal(0,sigma_subj);
  for (k in 1:N) { //
    mu = alpha +  beta_length * length[k] +
      beta_dependent * dependent[k] +
      beta_sentid * sentid[k] +
      beta_bid * bid[k] +
      beta_subjrate * subjrate[k] +
      gamma_subj[subjid[k]];
    time[k] ~ lognormal(mu,sigma);
  }
}
```

The results of OW (white paper) are presented in Table12. The results were same as the frequentist model results. In addition, the blmm evaluation was performed on OT and PB. The models of OT and PB did not converge. Further studies will investigate more sophisticated statistical models with the blmm.

## 8. Bibliographical References


Amano, S. and Kondo, T. (1998). Estimation of Mental Lexicon Size with Word Familiarity Database. In *Proceedings of International Conference on Spoken Language Processing*, volume 5, pages 2119–2122.

Asahara, M. and Kato, S. (2017). Between Reading Time and Syntactic / Semantic Categories. In *Proceedings of IJCNLP*, pages 404–412.

Asahara, M. and Matsumoto, Y. (2016). BCCWJ-DepPara: A syntactic annotation treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 49–58, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Asahara, M., Ono, H., and Miyamoto, E. T. (2016). Reading-Time Annotations for "Balanced Corpus of Contemporary Written Japanese". In *Proceedings of COLING*, pages 684–694.

Asahara, M. (2017). Between reading time and information structure. In *Proceedings of PACLIC*, pages 15–24.

Asahara, M. (2018a). Between reading time and clause boundaries in Japanese - wrap-up effect in a head-final language. In *Proceedings of PACLIC*, pages 19–27.

Asahara, M. (2018b). Between Reading Time and Zero Exophora in Japanese. In *Proceedings of READ2018*, pages 34–36.

Asahara, M. (2019). Word familiarity rate estimation using a Bayesian linear mixed model. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 6–14, Hong Kong, November. Association for Computational Linguistics.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical Introduction to Statistics using R*. Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67:1–48.

Cop, U., Dirix, N., Driegphe, D., and Duyck, W. (2017). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49:602–615.

Frank, S. L., Monsalve, I. F., Thompson, R. L., and Vigliocco, G. (2013). Reading-time data for evaluating broad-coverage models of english sentence processing. *Behavior Research Methods*, 45:1182–1190.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., and Fedorenko, E. (2018). The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., and Fedorenko, E. (2021). The natural stories corpus: A reading-time corpus of english texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1):63–77.

Hlavac, M. (2018). stargazer: Well-formatted regression and summary statistics tables. R package version 5.2.2.

Husain, S., Vasishth, S., and Srinivasan, N. (2015). Integration and prediction dificulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*, 8(3):1–12.

Kato, S., Asahara, M., and Yamazaki, M. (2018). Annotation of 'word list by semantic principles' labels for balanced corpus of contemporary written japanese. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information, and Computation (PACLIC 32)*.

Kennedy, A. and Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, 45(2):153–168.

Kliegl, R., Grabner, E., Rolfs, M., and Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1-2):262–284.

Kliegl, R., Nuthmann, A., and Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, 135(1):12–35.

Kuperman, V., Dambacher, M., Nuthmann, A., and Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology*, 63:1838–1857.

Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., and Kliegl, R. (2019). Russian sentence corpus: Benchmark measures of eye

movements in reading in russian. *Behavior Research Methods*, 51(3):1161–1178.

Luke, S. G. and Christianson, K. (2018). The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50:826–833.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 48:345–371.

Matsumoto, S., Asahara, M., and Arita, S. (2018). Japanese clause classification annotation on the balanced corpus of contemporary written japanese. In *Proceedings of the 13th Workshop on Asian Language Resources (ALR13)*, pages 1–8.

Miyauchi, T., Asahara, M., Nakagawa, N., and Kato, S. (2017). Annotation of information structure on the balanced corpus of contemporary written japanese. In *Proceedings of 2017 Conference of the Pacific Association for Computational Linguistics*.

NTT Communication Science Laboratories. (1999-2008). NTT Database Series [Nihongo-no Goitokusei]: Lexical Properties of Japanese.

Pan, J., Yan, M., Richter, E. M., Shu, H., and Kliegl, R. (2021). The beijing sentence corpus: A chinese sentence corpus with eye movement data and predictability norms. *Behavior Research Methods*.

R Core Team. (2020). R: A language and environment for statistical computing.

Sorensen, T., Hohenstein, S., and Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, 12:175–200.

Vasishth, S., Schad, D., Bürki, A., and Kliegl, R. (2021). *Linear Mixed Models in Linguistics and Psychology: A Comprehensive Introduction*.

## 9. Language Resource References

Shigeaki Amano and Kimihisa Kondo. (1999). *Tangoshinmitsudo: NTT Database Series: Nihongo-no Goitokusei Volume 1*. Sanseido.

Shigeaki Amano and Kimihisa Kondo. (2008). *Tangoshinmitsudo Zouho: NTT Database Series: Nihongo-no Goitokusei Volume 9*. Sanseido.

Uschi Cop and Nicolas Dirix and Denis Driegphe and Wouter Duyck. (2007). *Ghent Eye-Tracking Corpus*. Ghent University.

Richard Futrell and Edward Gibson and Harry J. Tily and Idan Blank and Anastasia Vishnevetsky and Steven T. Piantadosi and Evelina Fedorenko. (2021). *Natural Stories Corpus*. github.

Lena Jäger and Thomas Kern and Patrick Haller. (2021). *Potsdam Textbook Corpus (PoTeC): Eye tracking data from experts and non-experts reading scientific texts*. University of Potsdam.

Alan Kennedy. (2003). *The Dundee Corpus [CD-ROM]*. University of Dundee, Psychology Department.

Steven Luke. (2017). *The Provo Corpus: A Large Eye-Tracking Corpus with Predictability Norms*. OSF.

NTT Communication Science Laboratories. (2021). *NTT*. NTT Goi Database.

Naoko Sakuma and Mutsuo Ijuin and Takao Fushimi and Masayuki Tanaka and Shigeaki Amano and Kimihisa Kondo. (2008). *Tangoshinzousei: NTT Database Series: Nihongo-no Goitokusei Volume 8*. Sanseido.